DOCUMENT RESUME

ED 306 279 TM 013 130

AUTHOR Ackerman, Terry A.
TITLE An Alternative Methodology for Creating Parallel Test Forms Using the IRT Information Function.
PUB DATE Mar 89
NOTE 25p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, March 30, 1989).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *College Entrance Examinations; *Computer Assisted Testing; Computer Software; Higher Education; *Latent Trait Theory; Mathematics Tests; *Test Construction
IDENTIFIERS ACT Assessment; *Information Function (Tests); *Parallel Test Forms; TESTGEN Computer Program

ABSTRACT

The purpose of this paper is to report results on the development of a new computer-assisted methodology for creating parallel test forms using the item response theory (IRT) information function. Recently, several researchers have approached test construction from a mathematical programming perspective. However, these procedures require formidable computations, particularly as more constraints (i.e., the number of forms and the number of content areas) are added. In the test construction methodology proposed in this study, items are sampled from an ordered domain of item information values according to differences between test information curves of forms that are in the process of being created and a target test information curve. This new heuristic procedure, which uses a program known as TESTGEN, is being examined by the American College Testing (ACT) Program, which develops six parallel forms for the ACT Assessment Program each year. Research has concentrated on the Mathematics Usage Test. Results appear to be quite promising. Two data tables, one flowchart, and six graphs are presented. (Author/TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED306279

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Terry A. Ackerman

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

An Alternative Methodology for Creating Parallel Test Forms Using the IRT Information Function

Terry A. Ackerman

The American College Testing Program

Paper presented at the 1989 NCME Annual Meeting
March 30, San Francisco

TM013130

An Alternative Methodology for Creating Parallel
Test Forms Using the IRT Information Function

The purpose of this paper is to report results on the development of a new methodology for creating parallel test forms using the item response theory (IRT) information function. Recently, several researchers have approached test construction from a mathematical programming perspective. However, these procedures require formidable computations, particularly as more constraints (i.e., number of forms, number of content areas) are added. In the test construction methodology proposed in this study items are sampled from an ordered domain of item information values according to differences between the test information curves of forms that are in the process of being created and a target test information curve. Results appear to be quite promising. Underlying issues and future directions are also discussed.

An Alternative Methodology for Creating Parallel Test Forms Using the IRT Information Function

Currently the American College Testing Program (ACT) is investigating the role of computers in test form construction. Several researchers elsewhere have suggested various algorithms for constructing parallel test forms (Boekkooi-Timminga, 1986; Theunissen, 1985). These algorithms primarily use zero-one linear programming techniques and make use of the item information function from item response theory (IRT). However, these algorithms require large amounts of computer time.

A new heuristic approach is being examined by ACT to use in parallel forms development on the ACT Assessment Program (AAP) test for which six parallel forms are constructed every year. It is the purpose of this paper to present this new heuristic approach and some promising results. These results are based upon studies conducted with the old AAP test specifications but are believed to be directly applicable to the enhanced AAP which is targeted for administration starting in the fall of 1989.

Advantages of using the computer for creating parallel forms would include reduction in personnel time and the capability of matching forms to target information function. Ideally, forms created by computer would require only a substantive review of the item's text before final forms were prepared. Another indirect benefit is that equating of different forms may be made easier because of increased similarity across forms.

At ACT, research concerning the construction of parallel test forms has focused primarily on the AAP's Mathematics Usage test. This is because the Mathematics Usage test contains only discrete items and response data from this test consistently have been found to provide good IRT calibration results.

Background

AAP Mathematics Usage Test

The AAP Mathematics Usage Test is 40 items long and contains items which measure six content areas. These content areas and the numbers of items on the form used in this study were Arithmetic and Algebraic Operations (abbreviated AAO, 4 items), Arithmetic and Algebraic Reasoning (AAR, 14 items), Geometry (G, 8 items), Intermediate Algebra (IA, 8 items), Number and Numeration Concepts, (NNS, 4 items) and Advanced Topics (AT, 2 items). Each year six new forms having the above content specifications are created by test development personnel and designed to meet not only the content specifications, but also rigid statistical standards which include classical measures such as target mean item p-values and biserial correlations.

Classical Test Theory vs. IRT

In traditional or classical test theory, the degree of measurement precision provided by a test is described in terms of the standard error of measurement or the test reliability. Both of these statistics are group-dependent and may or may not generalize to other groups. Usually these indices are aggregated over items and individuals and therefore refer to the group's ability distribution as a whole.

In IRT, if the underlying assumptions of the response process are satisfied (e.g., local independence), the standard error of the latent ability estimate, $\hat{\theta}$, can be determined conditionally at each level of ability. However, instead of interpreting measurement precision in terms of the standard error, IRT practitioners use the inverse of the standard error, called information.

The precision of estimating an ability with an individual item g, can be derived for each level of ability, θ_i, using the formula

$$I_g(\theta_i) = \frac{P'_g(\theta_i)}{P_g(\theta_i)Q_g(\theta_i)}$$

where

$P_g(\theta_i)$ is the probability of a correct response for θ_i given by the IRT logistic response model,

$Q_g(\theta_i)$ is $1 - P_g(\theta_i)$, and

$P'_g(\theta_i)$ is the first derivative of $P_g(\theta_i)$ with respect to θ.

The information function for a test is the sum of the individual item information functions. The property of determining independent item contributions is not present in classical measurement. If two test forms, A and B, are strictly parallel, the test information curves will have identical shapes, indicating that for each θ_i, the measurement precision of test A equals that of test B.

Lord (1977) suggested procedures to construct parallel test forms which take advantage of this summative feature of item information. The procedure can be summarized as follows

1) Describe the shape of the desired or target test information function.
2) Select items with information functions that "fill-up" the target curve.
3) After each item has been added to the test, calculate the new test information function for the selected items. Determine what other types of items are needed to fill up the entire area under the targeted curve over the entire range of ability.
4) Continue the process until the test information curve approximates the targeted curve to a specified degree.

Method

A computer program which would create test forms that are matched to a specified target information curve was developed by the author. This program, called TESTGEN, incorporates the procedures suggested by Lord (1977). A flow chart of TESTGEN outlining the basic steps used to create parallel forms is shown in Figure 1.

Insert Figure 1 about here

To run TESTGEN, the user must supply an item parameter file which contains each item's content classification and IRT item parameter estimates. Using this information, the program creates two temporary look-up files: one which contains the content classification and item information values for user selected ability points, and one in which the pool item numbers are ordered in terms of decreasing information value at each user-specified ability level.

After the two look-up files have been created, TESTGEN will prompt the user to select one of two possible test construction procedures. The first test construction option requires that the user indicate the number of parallel forms that are to be created and the number of items per form. The user must then supply the program with all of the item numbers for each form. The program will then output the test information values at selected ability levels for each form as well as the content information function values for each form. This option would only be used by a user who wanted specific items on each form. The results could then be used to determine how parallel the constructed forms really were.

With the second option, the user is again asked to enter the number of forms that are to be created, the number of content categories per form, and the number of items per content. The user must also specify a target information curve by entering information values at each of the previously specified ability points. The user can opt to select specific items but is not required to select any--the program will select all of the items. If the user chooses not to select any items, the program will randomly select an item for each form.

After the initial item selection, TESTGEN will compute the difference, at each of the specified ability levels, between the target information values and the current test information value for each test form. These differences are then used to prioritize the order in which the forms which will receive items from the current pool. The form having the largest difference will receive from the pool the most informative item at the ability level where the largest difference occurs. After an item is selected for a particular form, the content of the item is checked. If the form has this particular content area filled, the second most-informative item at the specified ability level will be selected. This process is repeated until an item has been assigned to each form. At this point the test information for each form is recomputed and the process is continued until the necessary number of items has been reached.

At this point the program computes the difference between each form's test information function and the information function for every other form at each of the specified ability levels. This process, performed in a subroutine called the EQUALIZER, minimizes the difference in information between all of the created forms. Forms which are most discrepant are then examined to see which content block of items could be "swapped" to minimize this difference. Thus the first step in this equalizing phase is to exchange entire blocks of

content. After this process has been completed, the forms are again checked for differences. If differences still remain, individual items are "swapped" or exchanged between forms until a user-specified level of difference exists between all forms at all specified ability levels.

In the development phase, the primary focus was to see if TESTGEN and IRT methodology could be used to facilitate the parallel form construction process for the AAP math test by having the program create forms to match content requirements and a specified target information function. Currently forms are created to match content specifications and targeted p- and biserial correlational values. This paper examines the similarities and differences between test forms created by TESTGEN to match an IRT target information function and test forms created by ACT test development staff using target classical indices.

The Item Pool

The pool from which TESTGEN created forms consisted of 600 items; 520 items were from 13 previously administered AAP Math Usage tests and 80 items were from the Collegiate Mathematics Placement Program (CMPP). All of the CMPP items were classified as AT items. All forms were calibrated to fit the three-parameter logistic model using the IRT calibration program LOGIST IV (Wyngersky, Barton, & Lord 1982). All item parameter estimates were placed on a common scale. A plot of the pool information function (the sum of all the item information functions) and the six content information functions are shown in Figure 2. The most informative group of items were in the G content area. Interestingly, each of the content areas peaked at different places along the ability scale: AAO (had a maximum value of 24.20 at θ = -.20), AAR (68.31, .70), G (56.27, 1.10), IA (50.79, .90), NNS (17.64, .90) and AT (35.00, 1.30).

Insert Figure 2 about here

In the construction of parallel test forms, it was always necessary to determine the relative quality of the items in each content area in the item pool. The plot of the individual cntent information functions could be misleading, in part because each content area has a different number of items. Thus a second analysis of the items was done. The ability value at which each item had maximum information was computed. A profile table was then created. In constructing parallel forms which meet rigid content specifications, it is necessary to determine at which ability level each item is providing the maximum information. This information "profile" is broken down for specific information ranges for each content and is displayed in Table 1. The NNS content area had the least informative items, while G had the most informative items. Items in the AAO content area peaked in a very narrow ability range, whereas the AAR area had items which had maximums as low as $\hat{\theta} = -2.5$ and as high as $\hat{\theta} = 3.0$.

Insert Table 1 about here

TESTGEN was programmed to create six forms, which matched the test information value of AAP Math Usage Form 26A. (Items from 26A were also included in the 600 item pool.) Different numbers of ability points were tried; however, the results reported in this paper are for 13 specified points. The ability levels and information values respectively for the target information curve were (-2.0, 1.1), (-1.6, 2.0), (-1.2, 3.3), (-.8, 5.4), (-.4, 8.3), (.0, 12.1), (.4, 17.1), (.8, 21.3), (1.2, 18.0), (1.6, 10.8), (2.0, 5.8), (2.4, 3.1), and (2.8, 1.7).

Comparison of the TESTGEN results with the traditional procedures employed at ACT was done in two ways. First, the test information functions of the six TESTGEN tests were compared graphically for six forms created for the AAP. The second method used was a statistical approach. For each constructed form the expected score distribution was calculated using the form's IRT item parameter estimates and a N(0, 1) population density. The first four moments of the expected score distributions and the predicted reliability of each form were computed.

Results

A plot of the test information curves for six AAP Math Usage forms, 25B, 25C, 25D, 25E, 25F and 26A, is shown in Figure 3. Items from each of these forms were part of the TESTGEN item pool. It should be remembered that each of these forms was created to match targeted classical statistics and not IRT information values.

Insert Figure 3 about here

A plot of the test information function for each of the six forms created by TESTGEN (without the EQUALIZER) are shown in Figure 4. It was hoped that the test information curves would be almost coincident, but because each test was created in reference to the target curve without comparison to the other test forms, the curves vary somewhat. That is, items were selected for each form at particular thetas because the distance between the target information curve and the current test information curve for that form was greatest. However, the ability points at which items were selected may have varied from test to test for each item. Thus the individual test information curves increased at different rates at each of the selected theta values.

Psychometrically, the author has called this effect, asyncronous information inflation.

Insert Figure 4 about here

The effect of a subroutine called the EQUALIZER can be seen in Figure 5. In Figure 5, the curves for each of the "equalized" six TESTGEN test information functions (shown in Figure 4) are displayed. Compared to the form information curves in Figure 3, the generated forms appeared to be more parallel. However, they all overshoot the target information curve.

Insert Figure 5 about here

The first four moments of the expected score distribution and the predicted coefficient α reliability estimates of the six TESTGEN forms and the six forms (Form 25B-26F) created by traditional methods are shown in Table 2. Results show that the TESTGEN forms had items that were, on the average, more discriminating and more difficult than the six forms created with 26A. The expected score distributions for the computer generated forms were all more positively skewed than their traditional counterparts.

Insert Table 2 about here

An additional plot of the expected raw score densities for the six AAP forms and six TESTGEN forms are displayed in Figure 6. The generated forms appeared to be slightly more similar, especially at the upper end of the ability scale.

Insert Figure 6 about here

Discussion and Conclusion

The results from this study appeared to be quite promising. It was interesting to see how similar the expected score distribution results were for forms created using two distinct methodologies. That is, traditionally AAP forms are constructed by test development personnel to match targeted p-values and biserial correlations and TESTGEN, a computer program, creates forms to match a target IRT information curve.

A better comparison might have been to have use the same item pool for both methods. The AAP forms were created from a pool of pretested items, while the TESTGEN forms were created from items which appeared on nationally administered forms. An information profile, similar to that shown in Table 1, could be used to indicate the information similarities and differences between the pretest pool and the TESTGEN pool.

In the development of TESTGEN and the EQUALIZER, several idiosyncrasies which occur when summing test information functions were noted. One peculiarity was that by selecting the most informative items at particular thetas, the target information curve frequently will be exceeded by the generated test curves before the specified number of items is reached. This occurs because the selection routine is always pulling out only the most informative items and not items with average or little information, as might be found in a typical test. This accelerated growth rate has been termed "selection infonoma."

Several approaches are being tried to rectify this problem. These include creating more forms than required, thus decreasing the information in any one form. A second approach being tried is similar to one used in adaptive testing to avoid item over-use. The procedure is called "4-3-2-1" approach. That is, each time an item is considered for selection, it would be considered with the next three less informative items and together the four items would have respectively a 40%, 30%, 20% and 10% probability of being chosen. It is thought that this approach would also lower the average information in each test form.

A second problem encountered in programming TESTGEN was that whenever an item is selected to minimize the difference between the current form's information and the target information value, the item is selected at a theta value at which the item's information value is not a maximum value. That is, the selected item causes the current test form information function to bulge at a different theta value than at the theta value for which the item was selected. The author refers to this problem as "item cellulite."

To resolve this problem, every time an item is selected, TESTGEN checks the form's new test information values at the other theta values to insure that the form's test information curve does not "bulge out" of acceptable limits. If it does, the program will select a new item.

Currently the TESTGEN procedures are being replicated to see how much effect the initial selection of items (which is random when the program selects all of the items) has ave on the parallelism of the test forms. It is suspected that most of the items selected in the creation of, say, six forms, would also be selected in subsequent "six-form-runs" because the program is always trying to select the most informative items.

Future directions include expanding the program to include passage data; determining which type of target curves shapes are reproducible (e.g., multimodal, uniform); and comparisons with zero-one linear programming approach.

References

Boekkooi-Timminga, E. (1986). Algorithms for the construction of parallel tests by zero-one programming. (Research report 86-3) Enschede, The Netherlands: University of Twente.

Boekkooi-Timminga, E. (1987). Simultaneous test construction by zero-one programming. Methodika, 1, in press.

Boomsma, Y. (1986). Item selection by mathematical programming. Enschede, The Netherlands: University of Twente.

Lord, F. M. (1977). Practical applications of item characteristic curve theory. Journal of Educational Measurement, 14, 117-138.

Theunissen, T. J. J. M. (1985). Binary programming and test design. Psychometrika, 50, 411-420.

Table 1

The number of items having information in specific ranges at selected thetas for each content area

Content	Value of Max Info	Theta												
		-3.0	-2.5	-2.0	-1.5	-1.0	-0.5	0.0	.5	1.0	1.5	2.0	2.5	3.0
NNS	< 1.0					2	7	7	9	20	3	2	1	
	1.0-2.0											1		
	> 2.													
AT	< 1.0						1	9	15	16	16	8	8	18
	1.0-2.0								2	4	4		1	2
	> 2.0											2		
AAR	< 1.0		2	1	4	9	31	38	39	28	13	1		1
	1.0-2.0								4	8	3			
	> 2.0													
AAO	< 1.0					8	20	12	6	1				
	1.0-2.0						3		2					
	>2.0													
IA	< 1.0				1	3	7	12	23	31	13	1		
	1.0-2.0								5	2	5	1		
	> 2.0													
G	< 1.0					1	7	8	18	27	12	5		
	1.0-2.0							2	6	10	4	2		
	> 2.0									1		1		

Table 2

The mean p-value, biserial correlation, coefficient and reliability and the first four moments of the expected score distribution for six TESTGEN created forms and six AAP math forms

Form	$\bar{p}$	$\bar{r}_{bis}$	$\bar{X}$	σ	Skewness	Kurtosis	α
1	.49	.60	19.50	8.98	.45	-.66	.91
2	.48	.61	19.05	9.24	.44	-.79	.91
3	.47	.62	18.96	9.27	.45	-.79	.92
4	.48	.61	19.07	9.11	.45	-.77	.91
5	.50	.62	20.15	8.93	.37	-.77	.91
6	.49	.62	19.71	9.09	.44	-.82	.91
25B	.50	.56	19.95	8.19	.32	-.78	.89
25C	.51	.57	20.24	8.52	.30	-.77	.90
25D	.50	.58	19.83	8.59	.35	-.76	.90
25E	.49	.62	19.66	9.19	.31	-.87	.91
25F	.51	.55	20.21	8.23	.28	-.79	.89
26A	.51	.58	20.40	8.70	.35	-.81	.90

Figure Captions

Figure 1. A flowchart for the computer program TESTGEN.

Figure 2. IRT information curves for each pool content area.

Figure 3. Test information curves for AAP Math Usage Forms 25B, 25C, 25D, 25E, 25F, and 26A.

Figure 4. Six test information for the six computer generated forms.

Figure 5. Test information curves for the six computer generated forms after the EQUALIZER was applied.

Figure 6. Expected raw score distributions for the AAP Math Usage Forms 25B, 25C, 25D, 25E, 25F, and 26A.

Figure 7. Expected raw score densities for the six forms created by TESTGEN.

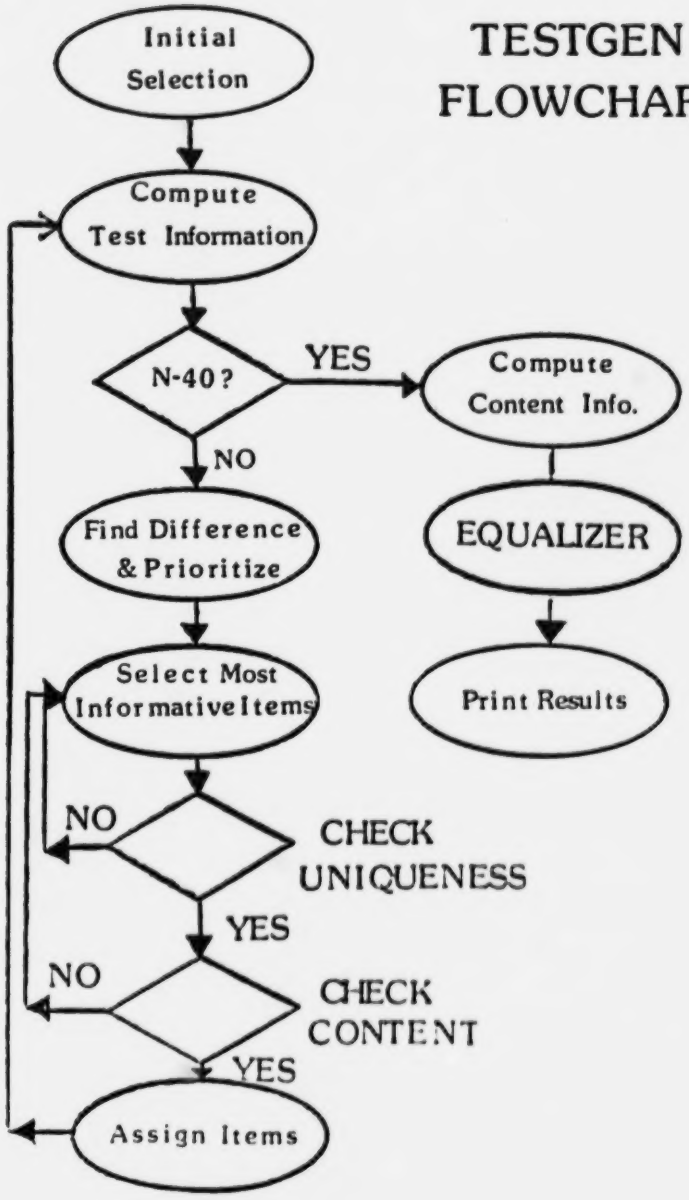

TESTGEN
FLOWCHART
Initial Selection
Compute Test Information
N-40?
YES
Compute Content Info.
NO
Find Difference & Prioritize
EQUALIZER
Select Most Informative Items
Print Results
NO
CHECK UNIQUENESS
YES
NO
CHECK CONTENT
YES
Assign Items

Figure 1

Figure 2

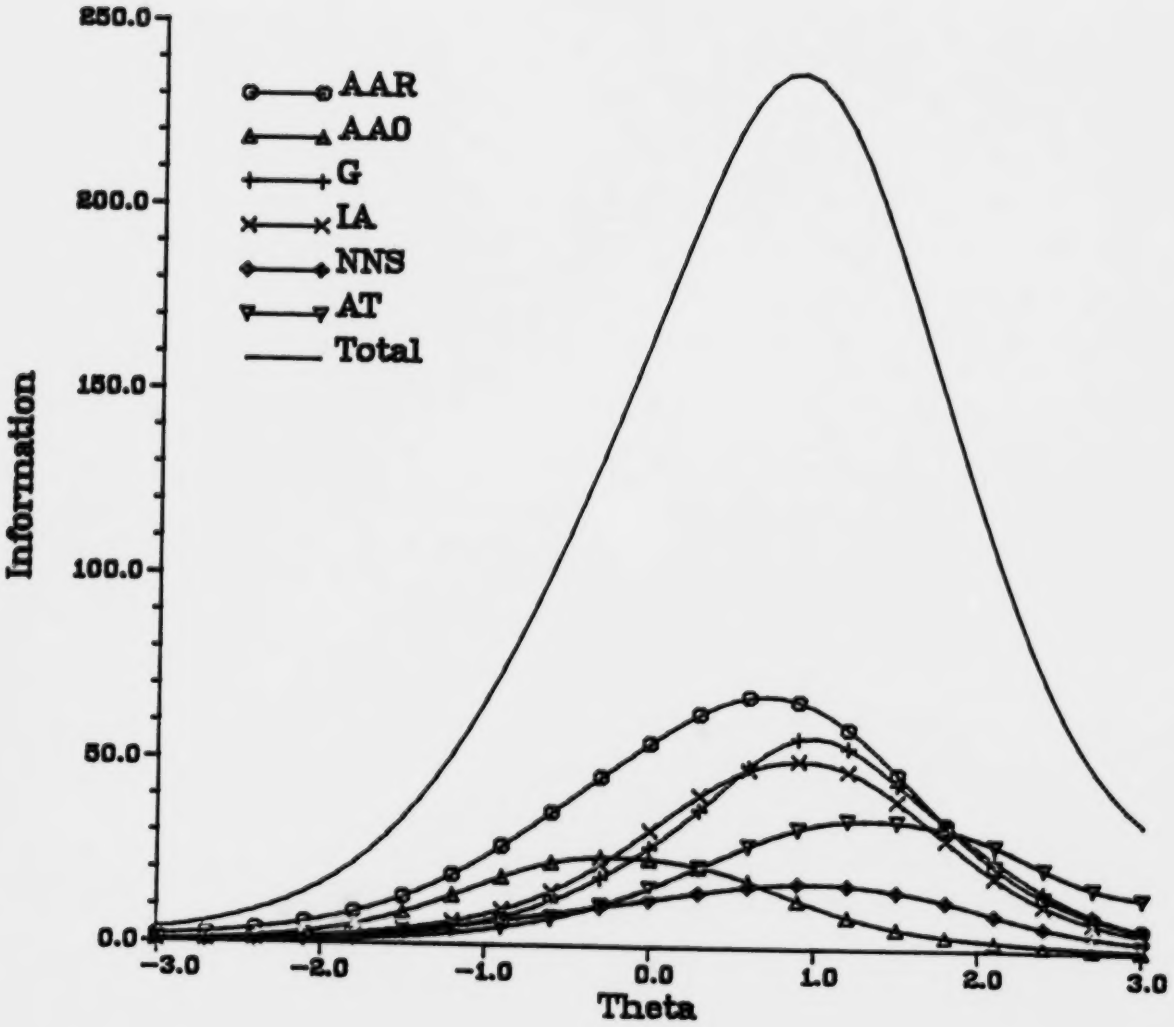

250.0
200.0
150.0
100.0
50.0
0.0
Information
AAR
AA0
G
IA
NNS
AT
Total
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
Theta

Figure 3

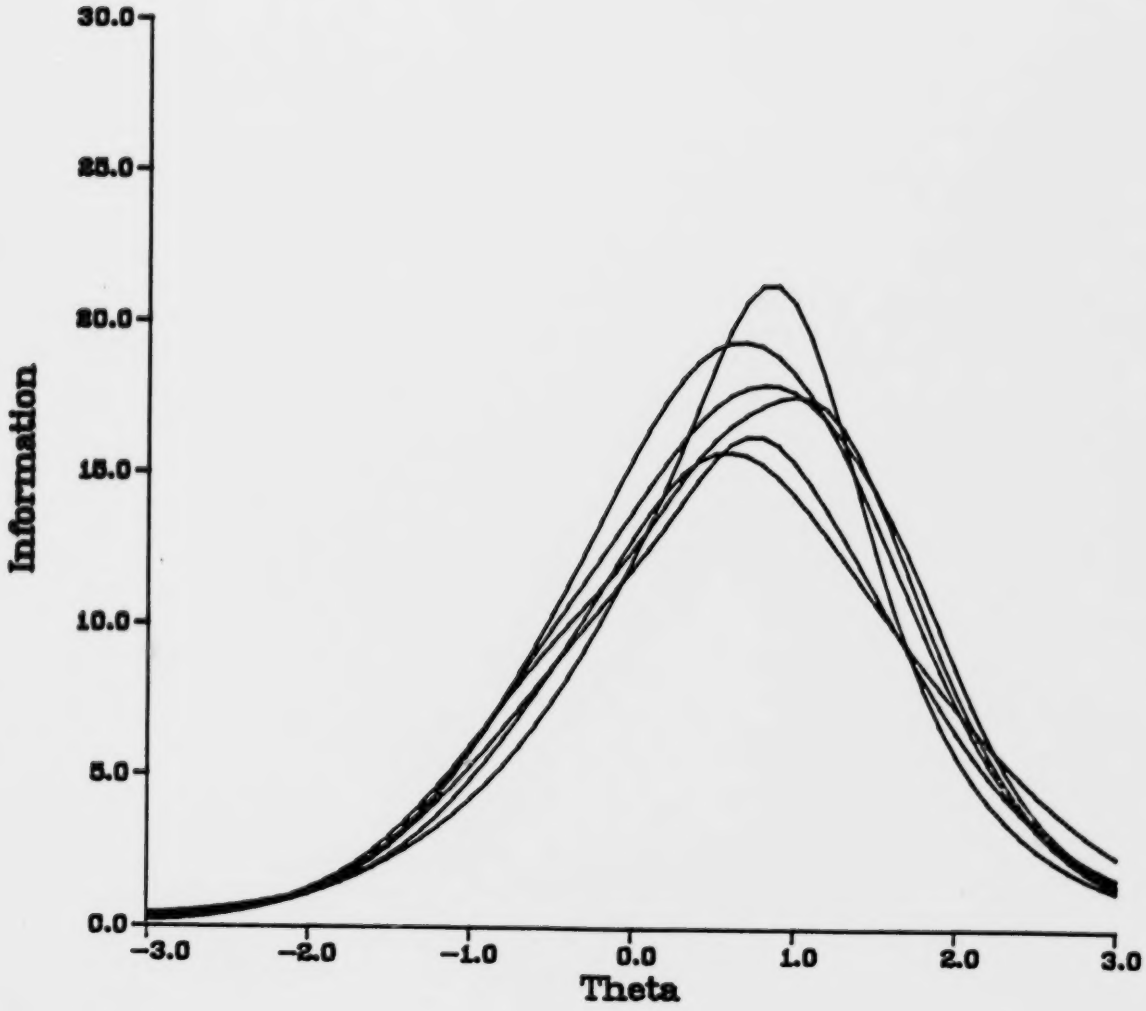


Figure 4

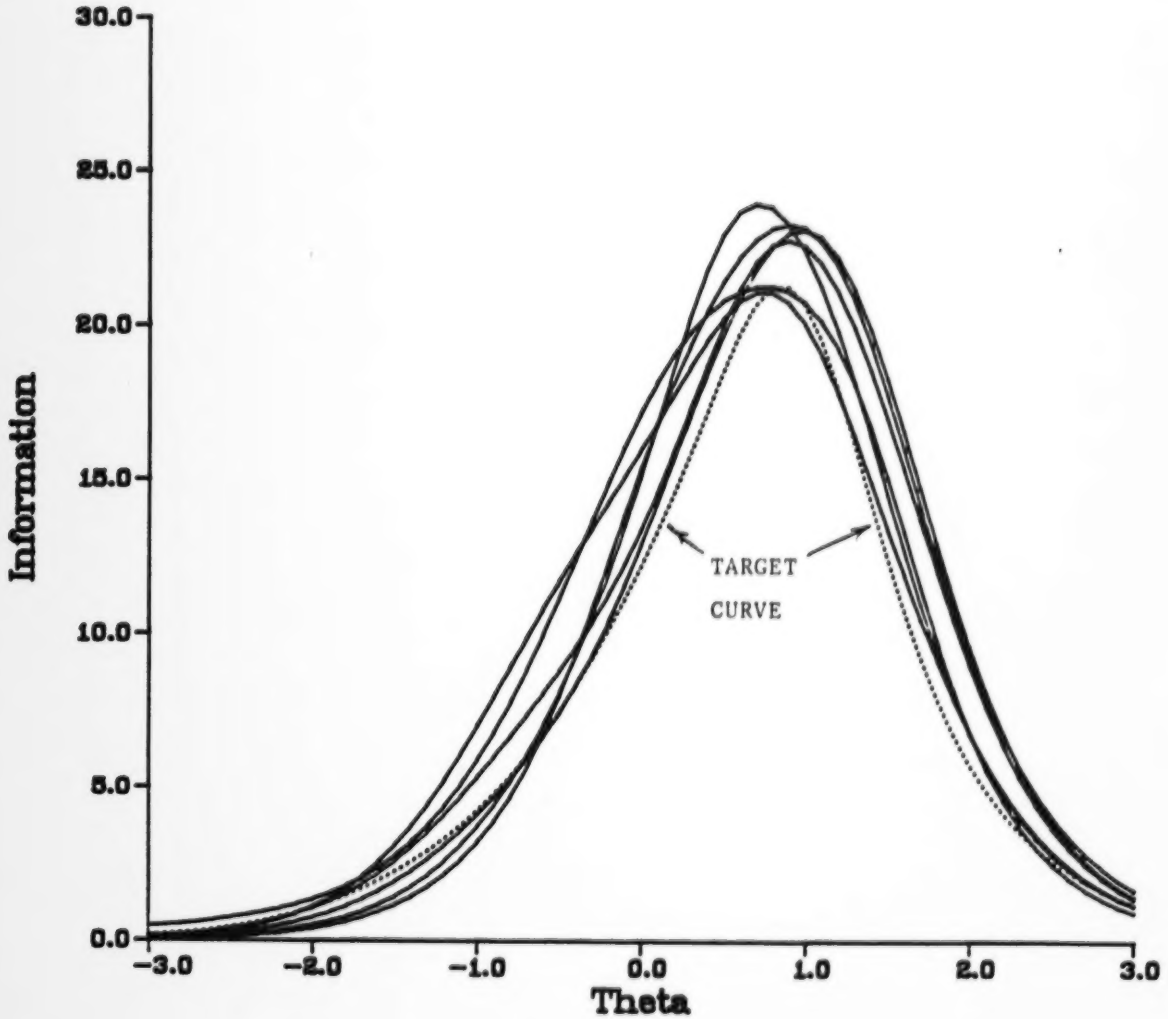

30.0
25.0
20.0
15.0
10.0
5.0
0.0
Information
TARGET
CURVE
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
Theta

Figure 5

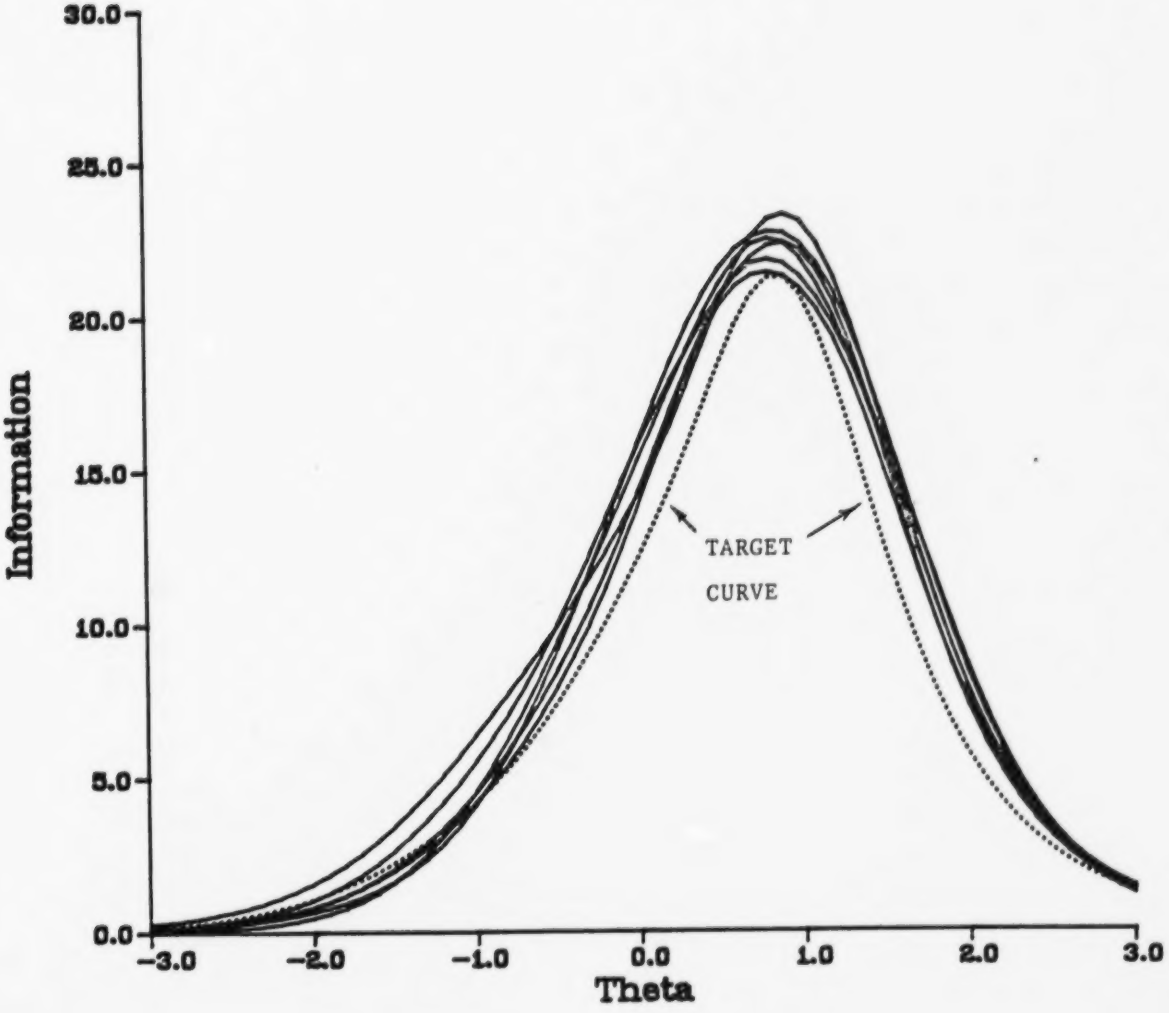
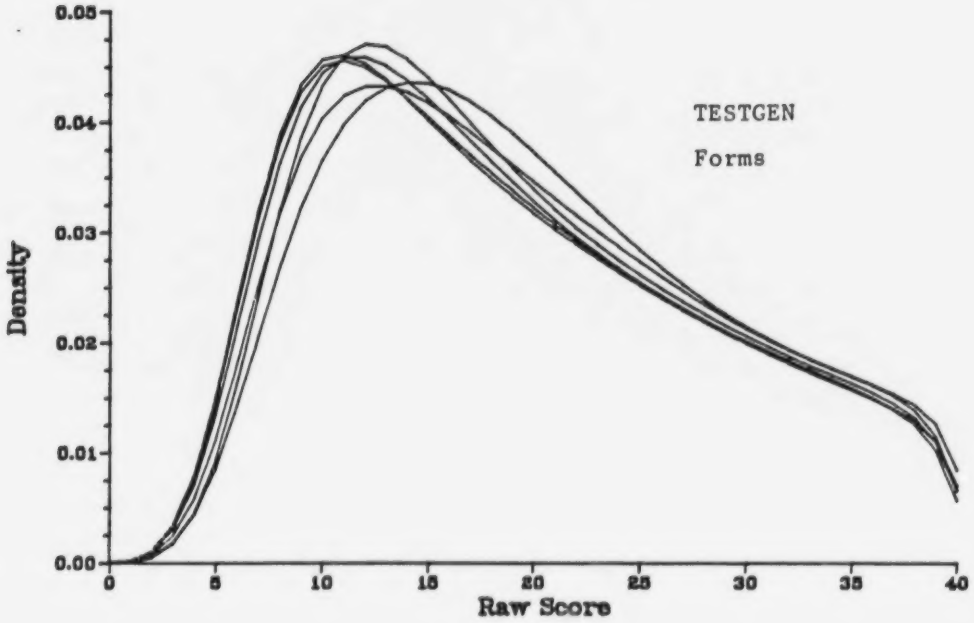
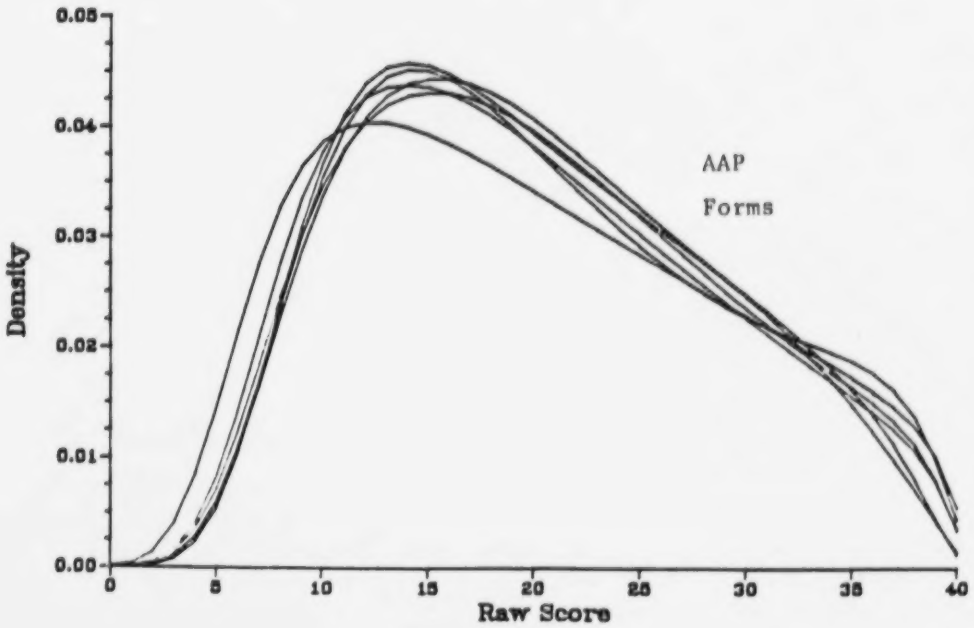

30.0
25.0
20.0
15.0
10.0
5.0
0.0
Information
TARGET
CURVE
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
Theta

Figure 6


0.05
0.04
0.03
0.02
0.01
0.00
Density
0
5
10
15
20
25
30
35
40
Raw Score
TESTGEN
Forms


0.05
0.04
0.03
0.02
0.01
0.00
Density
0
5
10
15
20
25
30
35
40
Raw Score
AAP
Forms